What is missing data and what should you do about it?*

Shiji Zhang

March 5, 2024

Table of contents

*Code and data are available at: https://github.com/Northboi/Missingdataessay

1 Introduction

In today's rapidly evolving information age, our lives are becoming increasingly intertwined with digital elements. People attempt to use data analysis to uncover patterns and mysteries in various aspects of life. Governments can formulate better policies by studying citizens' behaviors in certain areas. However, most statistical experiments involve some missing data, which not only affects our understanding of these experiments but can also introduce bias into the results if the missing data are not properly handled. This could lead to erroneous conclusions, potentially misleading society and individuals. Therefore, this article will demonstrate the importance of reducing missing data in statistical experiments and the potential threats they pose by further explaining what missing data are and how to handle them when encountered.

2 What is missing data

Firstly, missing data refers to the phenomenon where certain observations or variables in a dataset are not recorded or are left blank. There are many reasons for data to be missing. For example, in a statistical survey, participants may choose to leave a question blank due to concerns about revealing their personal privacy. This leads to data being missing. Generally, missing data types can be categorized into three categories: completely random missing, random missing, and non-random missing.(Little et al. 2014)

2.1 Missing completely at random

Here's an example: John and his teammates are conducting a statistical survey on academic stress at school. They distribute questionnaires to each student's desk in every class. The questionnaire includes variables such as age, academic performance, and study duration. However, John later discovers that not all students have filled out the questionnaire. Some students accidentally threw the questionnaire away, mistaking it for trash or scrap paper. Others may have lost the questionnaire on their way home. In this case, data are missing. However, since the missing data are not related to any variables or observations in the questionnaire, and students did not intentionally withhold information like age and academic performance, this type of missing data is purely due to random events. This type of missing data is called missing completely at random.(Little et al. 2014)

2.2 Missing at random

Similarly, let's understand what missing at random data means through an example. Suppose John and his team are conducting random surveys on people's intentions to purchase the new iPhone on the street. The survey questionnaire includes variables such as age and income.

Since income is considered a sensitive variable, some participants may choose not to disclose their income, resulting in missing data. In this case, the cause of the missing data is only related to the income variable and not to other variables such as age. Therefore, the missing data may be related to observed variables but not to variables that are not present in the questionnaire. This type of missing data is called missing at random.(Little et al. 2014)

2.3 Missing not at random

For instance, John and his team are currently researching the effectiveness of a certain medication in treating depression. Participants are required to provide information such as age and gender before the study begins. The entire statistical study will last for six months. At the beginning, participants are actively undergoing treatment and providing John with sufficient statistical data. However, halfway through the experiment, some participants lose contact with John, making it impossible to continue their participation in the study, resulting in missing data. In this case, the cause of the missing data is due to external factors affecting the participants, such as sudden family emergencies or other personal circumstances that lead to the interruption of the experiment. It is not related to factors already observed in the statistical survey, such as age and gender. Therefore, the probability of missing data in this case may depend on unobserved factors. This type of missing data is called missing not at random.(Little et al. 2014)

3 What should we do about missing data

3.1 Complete case analysis

Complete case analysis is a relatively simple but also hasty method of dealing with missing data. It involves directly ignoring the missing data in the dataset and only studying the parts with complete data. The advantage of this approach is its simplicity and lack of need for additional processing. However, the disadvantage is that it may reduce the sample size and overlook valuable information from participants with missing data, potentially leading to wasted information and unreliable results.

3.2 Imputation

Imputation is a method of handling missing values by using existing data (Baraldi and Enders 2010). Common imputation methods include mean imputation, median imputation, and model-based imputation. Mean and median imputation involve replacing missing values with the sample mean or median, respectively. Since the median is less influenced by outliers, using median imputation might be more robust in cases of uneven data distribution. Model-based imputation, on the other hand, constructs a regression model using information from other

variables to estimate missing values. Although this method is more complex, it often yields more accurate results.

3.3 Multiple imputation

Multiple imputation is a more complex and sophisticated method of handling missing data based on imputation (Baraldi and Enders 2010). It involves performing imputation multiple times to create several datasets with complete data. These datasets are then analyzed separately, and the analysis results are combined to obtain the desired outcome. Multiple imputation avoids wasting information and fully utilizes existing data. However, the downside is that multiple imputation is very complex and requires a considerable amount of time and effort.

4 Conclusion

In statistical experiments, the three types of data missingness — completely random, random, and non-random are quite common. Completely random missingness is unrelated to observed or unobserved variables. Random missingness may be related to one of the observed variables, while non-random missingness is related to unobserved variables. When faced with these types of missing data, it's essential to use appropriate methods for the most efficient missing data handling. Complete case analysis involves removing all observations with missing data, which can lead to significant errors in results. Imputation and multiple imputation seem to be more reliable methods. Imputation can estimate missing data using existing data through means, medians, or modeling. Multiple imputation involves multiple imputations to create several datasets with complete information, analyzing each dataset separately, and summarizing all results. Multiple imputation tends to be more accurate than imputation alone but requires more time and effort. Therefore, when dealing with different types of missing data, choosing the most effective handling method is crucial to ensuring the reliability of the final statistical research results.

[This essay was reviewed by my partner Zongcheng Cao, and the content of review can be found at https://github.com/Northboi/replication-of-impact-of-past-behaviour-normality-on-regret/issues and I have updated my essay based on his review.]

Reference

Baraldi, Amanda N, and Craig K Enders. 2010. "An Introduction to Modern Missing Data Analyses." *Journal of School Psychology* 48 (1): 5–37.

Little, Todd D, Terrence D Jorgensen, Kyle M Lang, and E Whitney G Moore. 2014. "On the Joys of Missing Data." *Journal of Pediatric Psychology* 39 (2): 151–62.